中国科学技术大学 计算机学院

2013 级研究生《机器学习与知识发现》期末考试试题

姓名：_____ 学号：_____ 成绩：_____

一. 数据对象间的相似性度量计算 （8 分）

(1) a=(1, 1, 0, 1, 1, 0), b=(0,1,1,0,1,0), c=(1,1,3,3), d=(3,3,1,1), 计算

i. a, b 的 Jaccard 相似系数（Jaccard Cofficient）

ii. c, d 向量空间余弦相似度（Cosine Similarity）

iii. c, d 的皮尔森相关系数（Pearson Correlation Cofficient）

解：$\mathrm{Jaccard}(a,b)=\frac{f_{11}}{f_{01}+f_{10}+f_{11}}=\frac{2}{1+2+2}=2/5$

$\cos(c,d)=\frac{cd}{\|c\|\|d\|}$=3/5

$\mathrm{corr}(c,d)=\frac{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\bar{y})^2}}=\frac{\frac{1}{3}(-4)}{4/3}=-1$

(2) 以上三种度量方法，哪些方法比较适合度量文本数据的相似性？ 简要说明。

答： Jaccard 和 Cosine。。。

二. 考虑下图的决策树。假设产生决策树的数据集包含 16 个二元属性三个分类 C_1、C_2 和 C_3。根据最小描述长度原则（MDL）计算每棵决策树的总描述长度。

- 树的整体描述长度由下式给出：

  Cost(tree, data) = Cost(tree) + Cost(data | tree)

- 树的每个内部节点用划分属性的 ID 进行编码。如果有 m 个属性，为每个属性编码的代价是 $\log_2(m)$个二进位。
- 每个叶节点使用与之相关联的类的 ID 编码。如果有 k 个类，为每个类编码的代价是 $\log_2(k)$个二进位。
- Cost(tree) 是对决策树的所有结点编码的开销。为了简化计算，可以假设决策树的总开销是对每个内部结点和叶结点编码开销的总和。
- Cost(data|tree) 是对决策树在训练集上分类错误编码的开销。每个错误用 $\log_2(n)$个二进位编码，其中 n 是训练实例的总数。

根据 MDL 原则，哪棵决策树更好？（10 分）

决策树 a 的总代价：2× 4+3× 2+7× log2 n = 14+7 log2 n.

决策树 b 的总代价： 4× 4+5× 2+4×log2 n = 26+4 log2 n.

根据 MDL 原则，
若 n<16。 树 a 好；
若 n>16. 树 b 好；
若 N=16，一样好
.

三. 请简述构建组合（集成）分类器的几种方法，并说明集成分类器能够改善分类器性能的原因。（12 分）

书上内容

四. 如下表数据，前四列是天气情况（阴晴 outlook，气温 temperature，湿度 humidity，风 windy）；最后一列是类标签，表示根据天气情况是否出去玩。（14 分）

(1) 根据上述训练数据，基于信息增益决策树应该选哪个属性作为第一个分类属性？

(2) 请画出两层决策树模型。

(3) 使用朴素贝叶斯方法预测测试样本（outlook=rainy, temperature=cool, humidity=normal, windy=FALSE）的类标号。
P(no|x)=2/5 * 1/4 * 1/4 *1/4 *1/2 = 1/320
P(yes|x)=3/5 * 3/6* 3/6 *4/6*5/6= 1/4
预测为 yes

(4) 对于含有连续型属性的样本数据，决策树和朴素贝叶斯分类能有哪些处理方法？离散化（等区间离散、等数据离散），高斯分布模拟

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes

五. 考虑下表中显示的购物篮事务的例子。（12 分）

事务 ID	购买项
1	{B，C}

2	{R，D，M}
3	{M，T，C}
4	{M，B，T}
5	{B，C，T}
6	{M，T，R，D}
7	{R，D，T}
8	{B，T}
9	{R，D，C}
10	{M，T，R，D}

(1) 从这些数据中，能够提取出的关联规则的最大数量是多少？（包括零支持度的规则）$\sum_{k=2}^{6}\binom{6}{k}(2^k-2)=602$

(2) 能投提取的频繁项集的最大长度是多少？（假定最小支持度>0）4

(3) 写出从该数据集中能够提取的 3-项集的最大数量的表达式。$\binom{6}{3}=20$

(4) 找出一个具有最大支持度的项集（长度为 2 或更大）。(D,R)

(5) 找出一对项 a 和 b，使得规则 $\{a\}\rightarrow\{b\}$ 和 $\{b\}\rightarrow\{a\}$ 具有相同的置信度。

(B,C) or (D,R)

课后习题：6.6 ，修改了数据集中的相应数据：BBread 与 RBeer 交换；TButter 与 DDiapers 交换。

六. 聚类分析 （14 分）

(1) 在聚类分析中，传统的 K-means 算法都有哪些局限性？有哪些相应的改进方法？

(2) 请简要描述聚类与关联分析的主要相似点和不同点。

(3) 请举出一个采用聚类作为主要的数据挖掘方法的实际应用例子。

七. SVM 计算题 （12 分）

已知正例点 $x_1=(1,2)^T$，$x_2=(2,3)^T$ ，负例点 $x_3=(2,1)^T$，试求最大间隔分离超平面和分类决策函数，并在图上画出分离超平面、间隔边界及支持向量。

拉格朗日乘子：(3,2,5). W1=-3, w2=7, b=-6. ???在算一遍？？

八. 优酷网是国内领先的视频分享网站，兼具影视、综艺、资讯三大内容形态。电视剧因其独特的表现形式，一直备受大众欢迎。优酷网提供了专门的电视剧播放频道，用户可以进入页面搜索浏览自己喜欢看的电视剧并下载收藏。目前，新剧《神探夏洛克 第三季》已在优酷网播放若干集，部门经理希望通过预测后续剧集的点击率，决定是否继续引进该剧。（18 分）

已知的数据包含很多部已上映完的电视剧的基本信息和播放记录，具体如下：

- 基本信息包括：电视剧名称、类型（剧情、警匪、悬疑…）、上映时间、导演、演员、集数、集均播放量、评论数、收藏数、豆瓣评分
- 播放记录指电视剧上映后每集每天的播放量，如下图所示，假设每部电

视剧每天只更新一集，Vji 代表第 j 集上映后第 i 天的播放量

	时间										
集序号	1	2	3	4	5	…	i	i+1	…	i+k	…
1	v11	v12	v13	v14	v15	…	v1i		…		
2		v21	v22	v23	v24	…	…	…			
3			v31	v32	v33	…	…	…			
4				v41	v42	…	…	…			
5					v51	…	…	…			
…						…	…	…			
j-1							vj-1,	…	…		
j								?	?	?	?

我们的目标就是：如上图所示，给定当前电视剧前 j-1 集的第 1 至第 i 天的播放记录（例如 j=11，i=7），预测第 j 集上映后每天的播放量即 $V_{j,i+1}$。

请利用你所学过的机器学习和数据挖掘的方法帮助部门经理解决该问题。

(1) 你认为上述已知的数据中，哪些特征或数据对于预测新一集的播放量有明显作用？

(2) 结合(1)中选择的特征或数据，使用什么技术、什么算法来解决预测问题，请具体说明之。

(3) 如果想预测一部新电视剧（所有集都没有上映）第一集上映当天的播放量，请简要给出一个解决方案。

中国科学技术大学 计算机学院

2016 级研究生《机器学习与知识发现》期末考试试题

姓名：_____ 学号：_____ 成绩：_____

一. 请画出训练误差和测试误差随模型复杂度变化而变化的曲线图，并解释 Ockham’s Razor（奥卡姆剃刀原则）。（5 分）

二. (朴素贝叶斯)试由下表的训练数据学习一个朴素贝叶斯分类器并确定 x = (2，S，T)的类判别结果 y。表中 $X^{(1)}$,$X^{(2)}$，$X^{(3)}$ 为特征，Y 为类标记。（10 分）

	1	2	3	4	5	6	7	8	9	10
$X^{(1)}$	1	1	1	2	2	1	2	2	3	3
$X^{(2)}$	S	M	M	S	S	L	M	M	L	S
$X^{(3)}$	T	T	F	F	F	T	F	T	T	F
Y	-1	-1	1	1	-1	-1	-1	1	1	1

三. (决策树)如下表数据，前四列是天气情况(阴晴 outlook，气温 temperature，湿度 humidity，风 windy)，最后一列是类标签，表示根据天气情况是否出去玩。（15 分）

(1) “信息熵”是度量样本集合纯度最常用的一种指标，假定当前样本集合 D 中第 k 类样本所占的比例为 （k=1，2，…，K），请问当什么条件下，D 的信息熵 Ent(D)取得最大，最大值为多少？

(2) 根据表中训练数据，基于信息增益决策树应该选哪个属性作为第一个分类属性？

(3) 对于含有连续型属性的样本数据，决策树和朴素贝叶斯分类能有哪些处理方法？

(4) 在分类算法的评价指标中，recall 和 precision 分别是什么含义？

(5) 若一批数据中有 3 个属性特征，2 个类标记，则最多可能有多少种不同的决策树？（不同决策树指同一个样本在两个决策树下可能得到不同的类标记）

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
rainy	cool	normal	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
overcast	cool	normal	TRUE	yes
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
overcast	hot	high	FALSE	yes

四、(SVM) 已知正例点 $x_1 = (1, 2)^T$，$x_2 = (2, 4)^T$，负例点 $x_3 = (2, 1)^T$，试用 SVM 对其进行分类，求最大间隔分离超平面，并指出所有的支持向量。（10 分）

五、(关联规则)考虑下表的购物篮事务：（12 分）

事务 ID	购买项
1	{牛奶，啤酒，尿布}
2	{面包，黄油，牛奶}
3	{牛奶，尿布，饼干}
4	{面包，黄油，啤酒}
5	{啤酒，饼干，尿布}
6	{牛奶，尿布，面包，黄油}
7	{面包，黄油，尿布}
8	{啤酒，尿布}
9	{牛奶，尿布，面包，啤酒}
10	{啤酒，饼干}

(1) 从这些数据中，能够提取出的关联规则的最大数量是多少（包括零支持度的规则）？

(2) 能够提取的频繁项集的最大长度是多少（假定最小支持度>0）？

(3) 写出从该数据及中能够提取的 3-项集的最大数量的表达式。

(4) 找出具有最大支持度的项集（长度为 2 或更大）。

六. (集成学习)回答以下问题（10 分）

(1) 试析随机森林为何比决策树 Bagging 集成的训练速度更快？

(2) 集成学习中多样性增强的方法有哪些？分别阐述这些方法适用的前提。

七. (聚类分析)回答以下问题（14 分）

(1) 领导者算法用一个点（称作领导者）代表一个簇，并将每个点指派到最近的领导者对应的簇，除非距离大于用户指定的阈值。在那样的情况下，改点成为一个新簇的领导者。与 K 均值相比，领导者算法的优点和缺点是什么？

(2) 使用如下表中的相似度矩阵进行单链和全链层次聚类。绘制树状图显示结果，树状图应清楚地显示合并的次序。

	P1	P2	P3	P4	P5
P1	1.00	0.10	0.41	0.25	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.85
P4	0.25	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

八．(降维)回答以下问题（8 分）

(1) 采用 PCA 方法降维后，有部分信息被舍弃了，简述为什么舍弃这部分信息是必要的？

(2) 对于高维空间到低维空间的函数映射不是线性关系的情况，可以采用什么方法进行降维？请阐述其核心思想。

九．综合题（16 分）

口碑是阿里巴巴集团与蚂蚁金服集团整合双方资源，联手打造的一家互联网本地生活服务平台。与淘宝、天猫不同，口碑面向的是线下实体商家。在推广口碑的过程中，我们遇到了一些商家推荐上的困难，因此需要你的帮助。

现有数据如下表：

数据集名称	具体数据项
在线用户行为	用户 id，线上商家 id，购买商品 id，商品类别，在线动作（0 表示点击，1 表示购买），行为时间
用户线下行为	用户 id，线下商家 id，地理位置，行为时间
商家信息	商家 id，商家可接纳的最高人数，商家地理位置

我们需要在特定的位置给用户推荐线下商家，商家个数可以是多个。由于口碑的用户使用记录较少，这给我们的商家推荐工作带来了一定的困难。而阿里巴巴拥有丰富的淘宝、天猫用户在线行为记录，我们希望能通过用户的在线行为数据来帮助预测其线下行为，改善线下推荐效果。

(1) 对用户的线下商家推荐，可以使用哪些机器学习算法？

(2) 你觉得用户的在线行为数据对线下行为是否能帮助线下行为的预测（绝大部分口碑用户有对应的线上行为记录）？请阐述理由。如果可以，请简要给出方案。

(3) 你认为哪些数据对推荐结果有较大影响？